PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68

A1

(11) International Publication Number: **WO 99/11822**

(43) International Publication Date: 11 March 1999 (11.03.99)

(21) International Application Number: PCT/US98/18261

(22) International Filing Date: 3 September 1998 (03.09.98)

(30) Priority Data:
60/057,479 3 September 1997 (03.09.97) US

(71) Applicant *(for all designated States except US)*: GENE LOGIC INC. [US/US]; 708 Quince Orchard Road, Gaithersburg, MD 20878 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: EVANS, Steven [US/US]; 418 North 38th Street, Omaha, NE 68131 (US).

(74) Agents: JONDLE, Robert, J. et al.; Rothwell, Figg, Ernst & Kurz, Suite 701 East, 555 13th Street N.W., Columbia Square, Washington, DC 20004 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD FOR PREDICTING PROGNOSIS AND TREATMENT RESPONSE FOR GENETICALLY BASED ABNORMALITIES

(57) Abstract

The present invention relates to a method for predicting prognosis and treatment response for genetically based abnormalities. The method utilizes clinically observable, collectible data from selected patient populations with diseases arising from genetic abnormalities, and uses such information to differentiate classes of patients with regard to specific differences in their genetic makeup, which in turn then characterizes outcomes such as prognosis and treatment response.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

TITLE OF INVENTION

Method for Predicting Prognosis and Treatment Response for Genetically Based Abnormalities

BACKGROUND OF THE INVENTION

The present invention relates to a method by which one can uncover patient and disease differentiations based on novel combinations of clinical and genotypical data (e.g., involving specific exon mutations within specific genes) which differentiate the patient population in ways that significantly enhance the ability to assess the prognosis and select suitable modes of treatment for the disease. In the most simple and basic model of man and disease, a single etiology such as an inherited genetic abnormality which, as a result, leads to a specific clinical manifestation reflective of the abnormality. For each such disease and its unique etiology, perhaps a specific treatment is available, and an associated prognosis possible. In many cases, such a simple model of a medical disorder is the exception rather than the rule. However, in areas in which we have inadequate understanding, this model may still be the assumption with which physicians must work until further differentiation is possible.

The publications and other materials used herein to illuminate the background of the invention or provide additional details respecting the practice, are incorporated by reference, and for convenience are respectively grouped in the appended List of References.

SUMMARY OF THE INVENTION

The present invention relates to a method by which one can uncover patient and disease differentiations based on novel combinations of clinical and genotypical data which differentiate the patient population in ways that

significantly enhance the ability to assess the prognosis and select suitable modes of treatment for the d In the method of the present invention, the process begins with a table of patient data that characterizes patients' medical clinical history in terms of specific gene's exon mutations or important polymorphism that patients have been found to carry, plus other clinical data. isease.

Given a selection of clinical data including (a) some particular disease state and (b) one or more patient clinical data factors the first goal (stage I) is to characterize (i.e., predict) the selected patient clinical data for the particular disease state chosen in terms of the presence of specific characteristics of specific genes. The second goal (stage II) is to take these identified characterizations derived in stage I and use them to differentiate the patient population expressing the chosen disease state in terms of the prognosis of the particular disease and in the efficaciousness of specific treatment modalities.

In summary, the method of the present invention takes clinically observable, collectible data from selected patient populations with diseases arising from genetic abnormalities, and uses such information to differentiate classes of patients with regard to specific differences in their genetic makeup, which in turn then characterizes outcomes such as prognosis and treatment response.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow-chart showing a method of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

In the method of the present invention, the process begins with a table of patient data that characterizes patients' medical clinical history in terms of specific gene's exon mutations or important polymorphism that patients have been found to carry, plus other clinical data which include but are not limited to: 1) any medical diagnoses they have and medical diagnoses their pertinent relatives have, particularly including first and second degree relatives (i.e., father and mother, siblings, and children, aunts, uncles, and

grandparents on both sides of their family); 2) ages of onset of such diseases; 3) associated medical and other conditions such as presence of colon polyps or high blood pressure; 4) ethnic origin; 5) race; 6) sex; 7) habits such as smoking, exposure to toxic chemicals; 8) occupations carrying medical risks such as agricultural workers; and 9) other additional information which may be available or obtainable in a patient's and patient's relatives' medical chart or file. For example, additional information for each patient might also include any treatments received, the outcome(s) of the treatment(s), ongoing patient management issues such as "time until first re-occurrence of the disease," etc.

Given a selection of clinical data including (a) some particular disease state and (b) one or more patient clinical data factors (such as "a significantly earlier onset" of that specific disease than is typical of the general population), the first goal (stage I) is to characterize (i.e., predict) the selected patient clinical data for the particular disease state chosen in terms of the presence of specific characteristics of specific genes, for example, specific mutated exons and/or over-expression. The second goal (stage II) is to take these identified characterizations derived in stage I and use them to differentiate the patient population expressing the chosen disease state in terms of the prognosis of the particular disease and in the efficaciousness of specific treatment modalities.

In summary, the method of the present invention takes clinically observable, collectible data from selected patient populations with diseases arising from genetic abnormalities, and uses such information to differentiate classes of patients with regard to specific differences in their genetic makeup, which in turn then characterizes outcomes such as prognosis and treatment response.

DEFINITIONS

In order to provide an understanding of several of the terms used in the specification and claims, the following definitions are provided:

"Data Mining Software" is widely available in the marketplace from numerous vendors. The technique and methodology essentially analyzes a

set of information, presented typically in a matrix array, in which each row of the matrix represents an example or instance of the phenomenon of interest and each cell within the row (i.e., column of the matrix) represents some descriptor of the phenomenon. For example, each row may represent a patient, with the first column containing a code indicating which disease (if any) the patient has had, the second column indicating what age he or she had the disease, the third indicating whether the patient is a smoker ("yes" or "no" value), etc. For certain rows (in our example, viewed as patients), these rows are marked "true" examples of any phenomenon of interest on the part of the user of the software (e.g., certain patients are identified as true cases of "hereditary breast cancer," or true cases of "patients who have survived a certain disease five or more years," or true cases of patients "who do not buy life insurance," etc.).

Data mining software uses computational techniques based on the mathematical theory of rough sets to select constituent elements (column values) for each row so that one or more boolean logical relationships among these elements characterizes (i.e., predicts) all the "true" examples that were marked, as noted above. One can select which column values the software may utilize to construct possible rules. For example, if the rows, viewed at patients, were marked for the "true" hereditary breast cancer cases and just exon mutations are specified as the allowable descriptive elements from the database, one boolean logical relationship (or "rule") the software could uncover might be that if there is a "yes" in the column marked "had a positive BRCA1 "gene test" or a "yes" in the column marked "had a positive BRCA2 gene test, " then any patient meeting this rule (A or B) is a member of the "true" set of hereditary breast cancer persons. If all medical factors in the database were allowed, another rule the software might develop could be: If a patient had beast cancer, the age of onset was before age 35, and there were two or more first degree relatives similarly affected, than the patient is one of the "true" hereditary breast cancer persons. Such a rule would be of the form (X and Y and Z). Through an efficient yet exhaustive process that is complete and systematic, data mining algorithms can find such valid rules or

patterns that can be constructed so that one or more of the "true" cases are characterized by that rule. The set of all such rules taken together define or characterize the set of "true" patterns provided, and as such, may be thought of as an expert rule-based system which defines or characterizes the patterned "true" set selected. Moreover, since the process examines the actual "true" cases to calculate the rules, its output can specify which cases are valid examples of which rules.

One off-the-shelf data mining product we have used in a part of our invention is DataLogic/R+1.5 from Reduct Systems, Inc., Regina, Saskatchewan, Canada, although numerous software choices and vendors are available.

"Disease" includes but is not necessarily restricted to any measurable condition of an individual that is not normal in some statistically probabilistic sense for some subset of the population, any malfunction of an individual or dysfunctional condition or genomic defect for a subset of the population, a genomic abnormality such as a mutation in a gene or identifiable polymorphism or other significant abnormality in part of the genome, or a genomic variation for some subset of the population which is measurably different from another subset, where such disease(s) can be manifest in behavioral changes, physical changes, tissue conditions, tissue or genomic-level conditions and changes, or other performance changes as is distinguishable from a different subset of the population.

"Exon Mutation" Particular parts of the DNA structure, the genes, provide the blueprints for the construction of proteins, assembling each protein in steps. Each step is dictated by and active coding region of the gene, and exon, which is interrupted by spans of non-coding regions (introns). If one of these coding regions has a flaw and can not properly assist in its role in the formation of the protein under construction, then the flawed area represents an exon mutation.

"Genotypical Data" includes but is not necessarily restricted to data about the genome of an individual, cellular factors or substances that influence or affect or alter the genome, measurable alterations in the genome,

data that reflect different levels of expression of the genome or parts of the genome or genome-related conditions, data that reflect the integrity or quality of the genome, or data that bear on the interrelationships between and among genomic elements of the cell or the cellular factors that influence the genomic elements.

"Patient Clinical Data Factors" includes but is not necessarily restricted to data that can measure or represent the status of an individual, any disease of that individual, or characteristics of that individual which might be included in a health-related database, with such data to also include information about an individual that distinguishes the individual from others (e.g., age, race) which may be used to characterize the healthy as well as the diseased individual.

"Prognosis" includes but is not necessarily restricted to possible outcomes, results, implications or consequences of disease, a subsequent state or status of a patient which could be psychological, behavioral, physiological, genomic-related, or other measurable difference as a consequence of disease or an abnormal condition.

"Treatment Outcome" includes but is not necessarily restricted to outcomes or results as a consequence of the administration of a drug or other therapeutic substance, or the administration of a therapeutic regimen which may involve ingesting any substance, a physical procedure, a behavioral procedure, a surgical procedure, any procedure that may lead to an alteration in the individual's genome or genomic-related condition, or any other health-related action intended to alter the health status of an individual. A treatment outcome may also be a measurable condition of the individual that distinguishes one state from a prior one even without knowing what agent created the transformation.

One method of the present invention includes the following steps:

Step 1 The first step is to select a particular syndrome, disease or combination of diseases of interest, such as hereditary breast-ovarian cancer. Data are assembled within a database as an array of patient information about patients who have exhibited

the selected disease or syndrome which includes for each patient, data regarding other specific diseases he or she may have had, diseases any first and second degree relatives have had as well as other immediate relatives, the age when any of the diseases occurred, any genetic mutations that have been determined, itemized by gene and by exons within the gene, and other clinical data which might be found in a clinical or medical record including treatments, drugs taken, outcome data with regard to treatment or the course of the disease for each patient as well as any relatives, etc.

Step 2 Second, one selects a particular disease, syndrome, or combination of diseases of interest d_i, such as hereditary breast-ovarian cancer.

Step 3a Next, one or more medical database elements (patient characterizations C_i) are chosen for patients with disease d_i, such as "had early onset of the disease" (for the example of hereditary breast-ovarian cancer this may be defined as the onset of the disease before the age of 40). The subset of selected patients who have the disease of interest and who exhibit these medical factor(s) chosen is defined to be the "true" examples (for the data mining software) of the pattern to be investigated. The subset of selected patients with the disease state but who do not exhibit the medical factors C_i is labeled as "false" examples of the pattern (for the data mining software).

Step 3b A similar set of patients with the selected disease is also chosen and the factor(s) chosen in the first step are negated ($-C_i$: they do not exhibit C_i), and the combination is considered the "true" patterns. In this example, patients with hereditary breast-ovarian cancer who were late-onset would be the "true" examples of the pattern. The subset of selected patients with the disease state but without exhibiting the negated factors is labeled as "false" examples of the pattern (for the data mining

software).

Step 4a/b Data mining methods are applied using as the "true" and "false" patterns those identified in steps 3a and 3b. The logical elements of the database which are specified for use by the data mining method (i.e., the specific columns for each data entry row) are the data about specific exon mutations of specific genes in individual patients. For example, assume that we have in our database entries for any detected exon mutations for the genes BRCA1, BRCA2, MLH1, and MSH2 for patients with hereditary breast-ovarian cancer. Then all of these are marked for the data mining process as the allowable elements (i.e., column values) which may be used to define the true set designated first in step 3a, then in step 3b.

Step 5a The output from the data mining software is a set of rules ("rule set") $R[C_i]$ that characterizes what logical mix of exon mutations of genes can be constructed so that the pattern is defined. These rules comprised of the occurrences of specific exon mutations "characterize" the designated "true" patterns. For our example, if one rule is "[Exon 2 of BRCA1 or Exon 3 of BRCA1]" arising in step 4a, then we know that if a patient exhibits the characteristics of having a mutation in either exon 2 or 3 of gene BRCA1, then that patient will be an example of the pattern of a "true" case of "hereditary breast-ovarian cancer of early onset."

Step 5b Output from data mining software for step 4b is a set of rules involving the exons for the negation of the medical condition under consideration, $R[-C_i]$, which is an analogous set of rules as arise in step 4a.

Step 6 The two sets of rules in steps 5a and 5b are compared (as described below) to see if there is any logical overlap of constituent elements (i.e., exon mutations) comprising the two rule sets. There constitutes a logical overlap if one condition, such as [mutation of exon 2 of BRCA1], occurs as an element in

both a rule in the rule set for 5a as well as in a rule in the rule set for 5b. If no such logical overlap exists for at least one exon, then the medical factor(s) (in our example, the condition "early onset" versus "late onset") is said to be a splitting condition for the disease based on one or more of the elements of definition (i.e., one or more exon mutations). The two sets of exon mutations arising in rules in steps 5a and 5b are the two splitting vectors (of exon mutations) for the splitting condition. The set of exons unique to each of the splitting vectors are called the splitting exons for the splitting vectors. It is ultimately such significant splitting conditions and their associated splitting vectors and splitting exon mutations for a disease state and associated medical factor(s) that we desire to obtain. This splitting condition means there are some differentiating clinical manifestation(s) for a disease state (in our example, "early" versus "late onset" for the disease hereditary breast-ovarian cancer) which can be reflected at the exon mutation level. This distinction of exon mutations in effect partitions the patient population with an otherwise apparently single umbrella disease (e.g., hereditary breast-ovarian cancer) into a potentially worthwhile, more refined division from a genetics point of view (such as, for example, "early onset hereditary breast-ovarian cancer" versus "late onset hereditary breast-ovarian cancer").

Step 6a If no significant splitting condition is found, another attribute is selected by the program from among the database attributes within the database (such as "there are two first degree relatives with the same disease"). This attribute is similarly tested for its prospects as a splitting condition. If one is obtained, then the second stage (step 7) of analysis is entered. If not, each database attribute can be tested in a loop, one at a time, then two attributes taken together can be selected, and in a loop, tested as a splitting condition. Again, if none is found, three

attributes taken together may be selected, etc. With a high speed computer, this cycling automatically considers all combinations in hopes of finding a splitting characteristic or a combination of attributes which will constitute a splitting condition.

Step 7 Once a splitting condition is obtained, (together with one or more associated combinations of splitting exon mutations comprising the associated splitting vector), a prognosis or treatment outcome attribute T_i and its negation $-T_i$ are selected where $T_i \neq C_i$. In our example, for the disease state "hereditary breast-ovarian cancer," and the attribute "early onset," selected treatment and outcome parameters (such as "five years or more of remission" and "received radiation treatment") can be identified. Generally, T_i can in fact be a boolean expression comprised of entries found in the database d_i.

Step 8a The first of the two set(s) of genetic exon mutations in the two splitting vectors from step 6 is used as the permissible characteristics to define the patterns for patient records which exhibit the prognosis or treatment outcome selected in step 7. For example, assume exons e_1, e_2, and e_3 in gene BRCA1 are the characterizing exon mutations (the splitting exons for the first vector) for the disease state "hereditary breast-ovarian cancer," and the splitting condition "early onset." Then the constituent elements e_1, e_2, and e_3 in the splitting vector represent the eligible characteristics that may be used with data mining software to define or characterize a prognosis or treatment outcome T_i for the disease condition and medical attribute(s) corresponding to the splitting condition (e.g., "early onset" and "hereditary breast-ovarian cancer"). As noted, data mining methodology is invoked, attempting to characterize these true patterns in terms of the exon set comprising the corresponding first splitting vectors. In our example, with the

exons e_1, e_2, and e_3 as the splitting set, this means one attempts to use these entries to characterize the "true" patients with their selected treatments and/or outcomes with boolean functions of the exon elements e_1, e_2, and e_3.

Step 8b As in step 8a, the exons in the first splitting vector is used to characterize the negative form of T_i.

Step 8c Analogous to step 8a, the exons in the second splitting vector are used to characterize the "true" set defined to be the patterns with d_i and $-T_i$.

Step 8d As in step 8c, the exons in the second splitting vector are used to characterize the "true" patterns defined to be the patterns with d_i and $-T_i$.

Step 9 If step 8a-d is successful, the "true" cases are characterized by rule sets from the data mining.

Step 10 If step 9 is successful for significantly characterizing the true set, then the corresponding constituent exons contained in each rule set are defined to be a successful two-stage splitting result (i.e., working for both stage I and stage II). One can derive for example the fact that if the patients who are early onset hereditary breast-ovarian cancer patients have certain exon mutations they are very responsive to radiation therapy, but late onset hereditary breast-ovarian cancer patients do not have such an exon mutation (but have their own particular set of exons) and thus are not responsive to radiation. It is noted that certain combination of results in step 10 are more desirable than others.

Step 11 To cycle through the database, one returns to step 7 and other prognosis and treatment outcomes attributes are selected (e.g., all hereditary breast-ovarian patients with one year reoccurrence rates, all with two year, all who responded favorably to radiation, all who responded favorably to chemotherapy, etc.) one by one using the first two sets of exons in $E_1[C_i]$ and $E_2[-C_i]$. If a more

than two-stage level is desired, we employ all four sets of exons derived in step 9 and enter step 7 with two pairs of exon sets rather than just two sets. Then in step 8, each pair is treated as each prior single set was treated. One may continue such a process for as many stages as desired. In the general case, the constituent elements of the splitting vectors derived are used as the permissible elements to define logical rules to characterize the prognosis condition or treatment outcome conditions selected. Once each prognostic measure or treatment outcome combination is selected these attributes may be combined into more complex patterns and tested, taking first two attributes in combination, then three, etc (e.g., does the exon set [e1, e2, and e3] characterize five year remission cases treated with both radiation and chemotherapy).

In selected cases, step 4a or 4b may be null, i.e., one-sided splitting conditions, obtaining no results, wherein 5a or 5b would be null. This is allowed since step 6 indicates that "*one* or both of [results] is not empty" (but it need not have been the case that both had to be "not empty"). The implications of this is that if only one "side" is non-empty, then that side is carried through in the remaining steps. Somewhat similarly, in step 7m, attribute T_i and $-T_i$ is selected, and results obtained if possible in the four parts of step 8. Again, if one or more parts of step 8 are not productive in deriving results, the process continues, and any output derived is exhibited in step 9. Finally, if no T_i is even selected at step 7, the process by default goes through steps 8 and 9 in the null case, and terminates as it loops back to step 7.

These above 11 steps create exon-based defining patterns derived from clinical manifestations which will characterize prognosis and treatment outcome patterns for differentiated versions of disease conditions. If achieved, a logical link has been uncovered connecting clinical attributes (such as early onset or late onset cancer) with a set of aberrant exons which in turn characterize medical prognosis or treatment outcomes.

A further method of the present invention includes the following steps, in addition to the steps of the above method:

In steps 3a and 3b of the method above, medical factors such as "early onset" are selected as interesting aspects of a medical disease (in this example, hereditary breast-ovarian cancer). Similarly in step 7, various prognosis or treatment outcomes are selected. In each case, exhaustive selection involving one or more combinations of factors may be chosen.

Step 1 We apply data mining to correlate mutational outcomes with patient attributes in order to guide choices for steps 3a, 3b and step 7. In our heuristic module's method for steps 3a and 3b, the "true" pattern of interest is the clinical family history of the patients who have (a) the selected disease state and (b) who have tested positive for any specific intragenic mutation(s) in any gene(s). All of the exons may be used by the data mining software to define the "true" set. All other patient records are labeled as "false." In step 7, at which point we have a splitting vector with available splitting exons, condition (b) becomes a positive occurrence of any one or more of these splitting exons.

Step 2 The data mining technology is then challenged to define what clinical medical attributes and what values for each attribute would define rules that characterized the disease state and mutations marked. This resulting set of rules might characterize the true set with such attributes such as "early onset" or "5 years of remission." All constituent elements comprising the rules heuristically suggest interesting choices for steps 3a, 3b or step 7 respectively.

The above additional two steps of the heuristic module yields choices for each of the two steps that may be more likely productive and informative than random selection alone.

EXAMPLES

The following examples are provided to further illustrate the present invention and are not intended to limit the invention beyond the limitations set

forth in the appended claims and amendments.

Example 1

Clinical data are collected on family histories of patients with the selected disease of hereditary breast-ovarian cancer who manifested a variety of mutations in the BRCA1 and BRCA2 genes (which yield breast and/or ovarian cancers). First, patients are defined who exhibited a 5382-insertion C mutation in BRCA1 as true. Patients who carried some mutation other than 5382-insertion C were labeled as false.

Using data mining technology to derive rules, the following results were obtained:

1. All rules contained the attribute "early age of onset" with very high values.
2. Nearly all rules indicated a rather intense family tree of breast cancers, with typically more than 3 cases within one generation of each other.

One of the actual rules takes the form of:

Ca32a=4 **AND** 5< Early <6 **AND** Gen=2

where *Ca32a* is the code for first degree relatives with breast cancer (which in this rule must have at least four such cases). *Early* refers to early onset of breast cancer, where three points are assigned if any case arises at or before the age of 35, two points are assigned if a case arises between age 36 and age 45, and one point is assigned for a case arising between 46 and 50. *Gen* refers to the presence of cancers in the same generation, which for this rule requires at least two cases in the same generation. Using our heuristic module, any of these three attributes (Ca32a, Early, and Gen) would be suitable as a choice in steps 3a and 3b.

Example 2

Gatekeeper genes directly regulate the growth of tumors by inhibiting growth or promoting death. Predisposed individuals who inherit one mutant copy of a gatekeeper gene need only one additional (somatic) mutation to initiate neoplasia. Moreover because the probability of acquiring a single somatic mutation is exponentially larger than the probability of acquiring two

such mutations, people with a gatekeeper gene mutation are not only at greater risk of developing tumors than the general population but since only one more "hit" is needed, they can be expected to express such tumors earlier in life (early onset).

Since we know that both BRCA1 and BRCA2 contain a region that act as a transcriptional-activation domain when it is fused to a DNA-binding domain from another gene (Milner, et al., 1997; Chapman, et al., 1996; Monteiro, et al., 1996), and transcription factors are often found among the gatekeeper class of cancer-susceptibility genes, this property indicates BRCA1 and BRCA2 may directly control cellular proliferation. Moreover BRCA1 can inhibit the growth of cells in which it is over-expressed (Holt, et al., 1996), and there is also a link between an inhibitor of cell-cycle-dependent kinase and BRCA1 protein (Hakem, et al., 1996). Thus there is reason to suspect BRCA genes of a gatekeeper function. From this gatekeeper status arises the expectation that BRCA mutations would lead to tumors much more frequently than the general population (i.e., higher risk), and also sooner than tumors arising usually (early onset). Both expectations are true for BRCA mutations.

However, Sharan et al. and Scully et al. provide evidence that BRCA genes are caretaker genes. Mutations in caretaker genes do not promote tumor initiation directly. Rather neoplasia occurs indirectly, since inactivation leads to genetic instabilities which result in increased mutations in all genes including gatekeeper genes, which then as noted above, may progress rapidly to tumor genesis. So if a patient inherited a caretaker gene defect, three somatic mutations would be needed to initiate cancer: mutation of the normal caretaker allele inherited from the unaffected parent, followed by mutations of both alleles of a gatekeeper gene. Moreover mutations in caretaker genes would not be expected to lead to sporadic cancers very often since four mutations would have to occur (two caretaker alleles plus two gatekeeper alleles). BRCA genes might be added to the caretaker list since mutations in BRCA1 and BRCA2 are rarely found in sporadic cancers. Thus we have two contending roles for BRCA1 and BRCA2, gatekeeper and caretaker.

Before proposing a resolution of this issue, we may take note of the fact that both BRCA genes bind to Rad51, a protein that is involved in maintaining the integrity of the genome (Sharan, et al., 1997; Scully, et al., 1997). Moreover Sharan et al. report that BRCA2 knockout mice show early embryonic lethality and hypersensitivity to irradiation, similar to that observed in RAD51 knockout mice. Earlier BRCA1 was found to bind to RAD51 (Scully, et al., 1997), and BRCA1 knockout mice also show early embryonic lethality (Hakem, et al., 1996; Liu, et al. 1996). We add the fact that studies in both yeast and mammalian cells indicate that RAD51 is involved in resolving double-stranded DNA breaks in recombination-linked repair, and thus we can expect that disruption of the BRCA/Rad51 pathway might be expected to lead to genetic instability. Consistent with this interpretation is the observed hypersensitivity to irradiation in embryonic and trophoblast cells from BRCA2 and Rad51 knockout mice, Sharan et al., 1997; Lim, et al., 1996.

Putting all of this together, we could postulate two classes of exon mutations in BRCA genes: (1) exon mutations related to gatekeeper functions but unrelated to genomic stability regions such as the Rad51 binding site would disrupt the gatekeeper status of the BRCA genes, yielding autosomal dominant early onset breast cancer, and (2) mutations unrelated to the gatekeeper status but harmful to genomic stability regions such as the Rad51 binding site disrupt the caretaker status of the BRCA genes, yielding late onset breast cancer. In summary, different exon mutations can affect either caretaker or gatekeeper-like functions in BRCA genes. Thus radiation of early onset breast cancer might be more efficacious since genomic instability arising from a compromised DNA caretaker role is not the primary issue, but conversely, radiation might be far less efficacious in late onset where the caretaker role of the gene and genomic instability (as in the knockout mice) is of significance.

Broadly speaking, this discussion demonstrates that there is a reasonable prospect that there is a connection to be uncovered between clinical features (such as early versus late onset) and exon positions and concomitant prognosis or treatment outcomes.

The exons involved in all BRCA1 patients who were 50 years or older, with no relatives of these patients under 40 were examined. The exons that characterized or predict these cases are: 5272; 300C; exon 5 missing

Then we looked at all BRCA1 patients who expressed cancer under the age of 40, and who had no relatives over 50 with breast cancer. The exons predicting this group which did not overlap in any way the first set were: 332 11T; 188del11; 4713intC.

Thus early versus late stage onset (as described by the example) constitutes a total splitting condition, and the exons found would distinguish early from late onset. Prognosis and treatment outcomes were then characterized from these two different sets of exon mutations, completing the application of the method. In this example, the prognosis condition "P" indicating high penetrance of the cancer (many generations), determined how the two sets of exons characterized P and -P. The output was 4 rule sets, which constitute the results of the process.

Example 3

Although specific exon mutations within specific genes may be utilized for the genotypical (i.e., genome-related) data, our method is not exclusively or restrictively constrained just to exon mutations, this merely being one type of genomic-related data, but rather genotypical data of other formulations as well.

Specification of disease in the method of the present invention may not be merely a traditional labeling such as pneumonia or diabetes but may in fact be dysfunctional conditions abnormal conditions as exhibited in tissue, or otherwise some abnormalities that can be contrasted with a baseline normality. The abnormality may even as yet have no formal name associated with it in medical science, and may be for example an over-expression of some gene for which medical science is not certain what to call or label the abnormality other than to contrast it to a baseline normal condition.

Concomitantly, a prognosis may include a specific outcome or condition or state for a patient, and also may include a level of physiological status (measured by laboratory values), or a genomic-related status such as a

measure of the degree of over-expression of a gene which can characterize an organism. As yet another variation, the outcomes in step 5a and 5b may not be necessarily so simple as the absence or presence of just an exon, but also may be the absence or presence of any factor used in the database which was stated to be comprised of clinical and genotypical data.

Moreover, it is also the case that the outcome in steps 5a and 5b may not be merely so simple as the presence or absence of a data element but also may be a function involving a data element such as element e1 is less than or equal to a certain value, greater than or equal to a certain value, or in general, any arithmetic function of the data element. Such variations of genotypical descriptions, disease states, prognosis states, and presence or value of a data element are all within the purview of the method's arena of application, although these variations are not inclusive of all variations, but only articulate some of the additional possible variations.

In this example, we describe such a variation of the type of genotypical data that can be used in applying this method, variations on the range of disease or abnormal situations to which the method may apply, variations on the prognosis states that may be selected in using this method, and variations in the data elements that comprise the set of rules from the data mining which characterize the data set.

To begin, specific over-expressed gene for a specific tissue is selected as the "abnormal" state (i.e., the "disease" for the method). As the medical factor Ci of interest, a specific range for the over-expression of the gene as the medical factor of interest is selected, with the specific range as measured by some laboratory method. The database includes the specific ranges of expression of this gene for each patient plus clinical data to describe patients, typical of data to be found in any comprehensive medical record. Thus with this selection, step 3a becomes: the true set are those patients with the over-expressed gene (their abnormal condition) with the value of the over-expression within some specifically designated range (as the medical factor of interest) along with their associated clinical and genotypical descriptors which may include exon mutation data, as well as also include other medical

descriptors as well.

In this example, step 3b constitutes the contrasting group of what amounts to those patients who do NOT have the same range of the gene over-expression, but otherwise are the same. Working with this example, we can derive by applying step 4a/b what characterizes the two different groups in step 3a and 3b. The items that characterize the two groups can be the presence or absence of descriptor factors or arithmetic measures of elements such as "patient's age is less than some value."

Thus one could use this approach when we have different levels of gene over-expression for some tissue set and we want to characterize this level of over-expression from the non-similar levels of over-expressors. The splitting condition becomes those elements that cleanly characterize one group from the other. If there exists a further outcome or prognosis as per step 7, the method steps are continued, however if none exists, we can, by default select none, default to step 9, which defaults to step 7 again under the condition "none remaining," which ends the process.

While the invention has been disclosed in this patent application by reference to the details of preferred embodiments of the invention, it is to be understood that the disclosure is intended in an illustrative rather than in a limiting sense, as it is contemplated that modifications will readily occur to those skilled in the art, within the spirit of the invention and the scope of the appended claims.

Example 4

The method of the present invention can also be applied where a disease d_i is a medical condition such as stroke, the clinical attribute C_i is a drug X is given to the patient, and its negation $-C_i$ is drug X not given (i.e., a placebo or null drug is given), and the rest of the characterizing elements are a mixture of clinical data and genotypical data. Steps 4a and 4b, 5a and 5b, and step 6 are then followed. In step 7 the outcome T_i selected is "a positive response to drug X as measured by the results of a trial for drug X" while its negation is "no positive response by the results of a trial for drug X." Hence step 8 characterizes the drug-taking positive responders, the placebo

positive-responders, and characterize the drug-taking non-responders and the placebo-taking non-responders.

Example 5

The method of the present invention can further be applied where a disease to be abnormal tissue of some type, the attributes C_i is a series of genes differentially expressed for the abnormal tissue, and then characterize the tissue in step 4 with the set of said genes in terms of clinical and genotypical data available for the tissue. In step 7, the treatment outcome can be measured in terms of some transformation of the gene series selected as C_i. Thus, step 8 obtains the data elements which characterize the abnormal tissue with a baseline genetic characterization that has undergone a genetic transformation, and that has failed to undergo a genetic transformation. One can also obtain the characterization of the tissue which did not exhibit the genetic series C_i but which underwent a transformation as well as did not undergo a transformation.

REFERENCES

1. Milner, J. et al. *Nature* **386**, 772-773 (1997)
2. Sharan et al. *Nature* **386**, 804-810 (1997)
3. Chapman, M.S. et al. *Nature* **382**, 678-679 (1996)
4. Monteiro, A.N. et al. *Proc. Natl Acad. Sci USA* **93**, 13595-13599 (1996)
5. Holt, J.T. et al. *Nature Genet.* **12**, 298-302 (1996)
6. Hakem, R. et al. *Cell* **85**, 1009-1023 (1996)
7. Scully, R. et al. *Cell* **88**, 265-275 (1997)
8. Liu, C.Y. et al. *Genes Dev.* **10**, 1835-1843 (1996)
9. Lim, D.S. et al. *Mol. Cell Biol.* **16**, 7133-7143 (1996)

CLAIMS

I claim:

1. A method for predicting the prognosis and a treatment response for genetically based abnormalities, comprising the steps of:

selecting a disease for analysis;

selecting a disease factor of said disease for analysis;

selecting a genetic abnormality of said disease for analysis;

marking said genetic abnormality as a first splitting factor if said genetic abnormality is existing in persons exhibiting said disease factor and not existing in persons not exhibiting said disease factor;

selecting a treatment option for said disease factor; and

marking said treatment option as a second splitting factor if a desired treatment outcome is existing in persons who received said treatment option and not existing in persons who did not receive said treatment option.

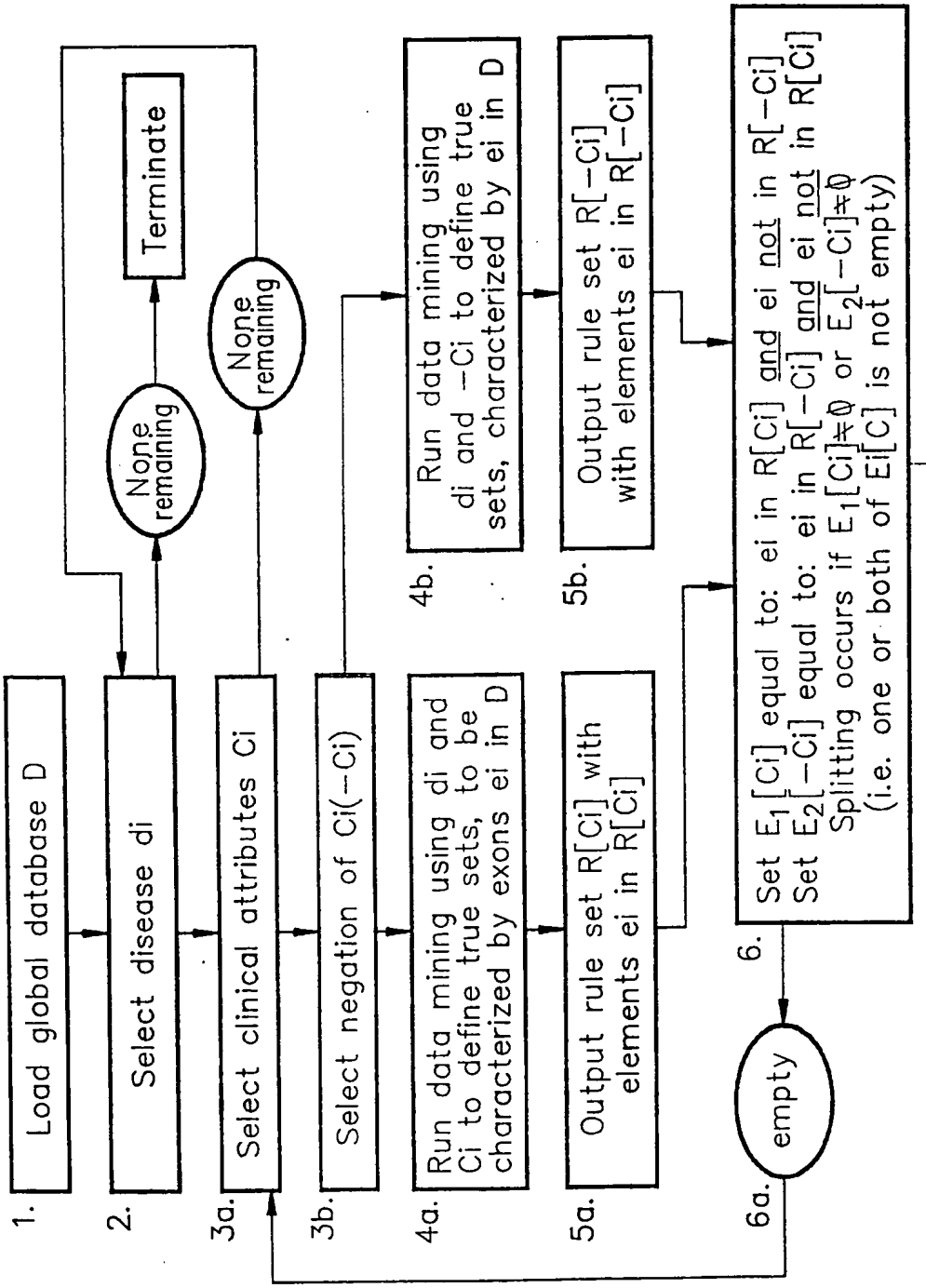
1. Load global database D
2. Select disease di
3a. Select clinical attributes Ci
3b. Select negation of Ci(−Ci)
4a. Run data mining using di and Ci to define true sets, to be characterized by exons ei in D
4b. Run data mining using di and −Ci to define true sets, characterized by ei in D
5a. Output rule set R[Ci] with elements ei in R[Ci]
5b. Output rule set R[−Ci] with elements ei in R[−Ci]
6. Set E1[Ci] equal to: ei in R[Ci] and ei not in R[−Ci]
Set E2[−Ci] equal to: ei in R[−Ci] and ei not in R[Ci]
Splitting occurs if E1[Ci]≠Φ or E2[−Ci]≠Φ
(i.e. one or both of Ei[C] is not empty)
6a. empty
None remaining
Terminate
None remaining
A

FIG. 1A

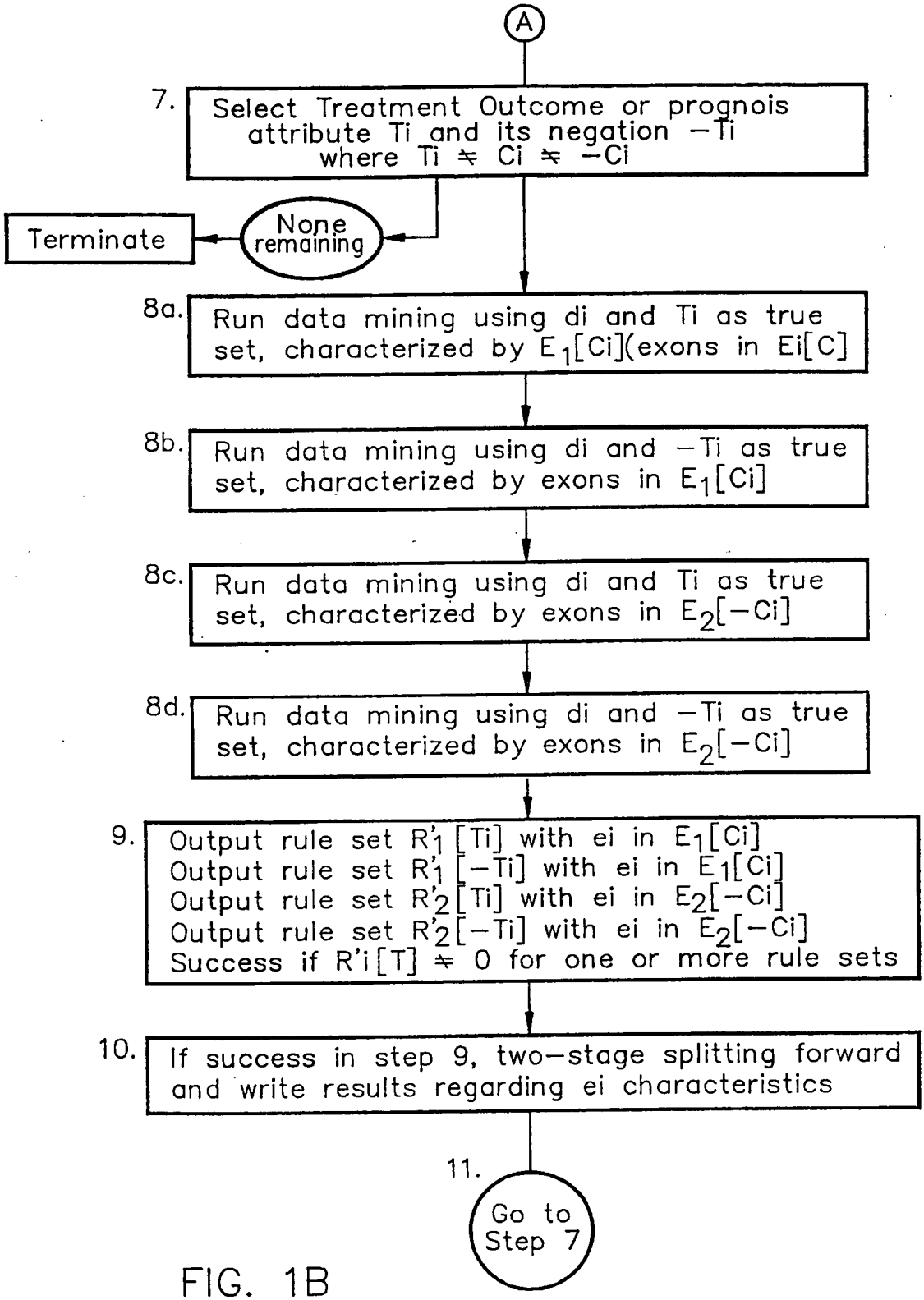
A
7.
Select Treatment Outcome or prognois attribute Ti and its negation −Ti where Ti ≠ Ci ≠ −Ci
None remaining
Terminate
8a.
Run data mining using di and Ti as true set, characterized by E1[Ci](exons in Ei[C]
8b.
Run data mining using di and −Ti as true set, characterized by exons in E1[Ci]
8c.
Run data mining using di and Ti as true set, characterized by exons in E2[−Ci]
8d.
Run data mining using di and −Ti as true set, characterized by exons in E2[−Ci]
9.
Output rule set R'1 [Ti] with ei in E1[Ci]
Output rule set R'1 [−Ti] with ei in E1[Ci]
Output rule set R'2[Ti] with ei in E2[−Ci]
Output rule set R'2[−Ti] with ei in E2[−Ci]
Success if R'i[T] ≠ 0 for one or more rule sets
10.
If success in step 9, two−stage splitting forward and write results regarding ei characteristics
11.
Go to Step 7

FIG. 1B

INTERNATIONAL SEARCH REPORT

Int tional Application No
PCT/US 98/18261

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MERAJVER ET AL: "Risk assesment and presymptomatic molecular diagnosis in hereditary breast cancer" CLINICS IN LABORATORY MEDICINE; vol. 1, no. 16, March 1996, page 139 139 XP002079197 see whole doc. esp. p.142, fig.2 and p.150 table 4 ---	1
X	GRANT S F A ET AL: "REDUCED BONE DENSITY AND OSTEOPOROSIS ASSOCIATED WITH A POLYMORPHICSP1 BINDING SITE IN THE COLLAGEN TYPE I ALPA 1 GENE" NATURE GENETICS. vol. 14, no. 2, October 1996, pages 203-205, XP002035047 see whole doc esp. table 1 ---	1
	-/--	

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 December 1998	18/12/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Müller, F

3

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Ir ational Application No
PCT/US 98/18261

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document. with indication.where appropriate. of the relevant passages	Relevant to claim No.
X	WO 97 06180 A (MEDICAL SCIENCE SYSTEMS INC ;KORNMAN KENNETH S (US); DUFF GORDON W) 20 February 1997 see whole doc. esp. claims and p.8 line 25 ff	1
Y	PIENTA K J: "RISK FACTORS FOR PROSTATE CANCER" MICHIGAN ONC. JOURNAL, vol. 6, 1995, pages 4-7, XP000617732 see the whole document	1
Y	WO 96 20288 A (CTRC RES FOUNDATION ;UNIV MICHIGAN (US)) 4 July 1996 see whole document, esp. page 3, line 20 ff. and claims	1
P,X	WO 97 43446 A (GEMINI INTERNATIONAL HOLDINGS ;RALSTON STUART HAMILTON (GB); GRANT) 20 November 1997 see whole doc. esp.p.9, 2. par. line 8 ff.	1

3

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int tional Application No
PCT/US 98/18261

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9706180 A	20-02-1997	US 5686246 A	11-11-1997
		AU 6763396 A	05-03-1997
		BR 9604044 A	09-06-1998
		CA 2228424 A	20-02-1997
		EP 0769560 A	23-04-1997
		EP 0853630 A	22-07-1998
		NO 980439 A	13-03-1998
WO 9620288 A	04-07-1996	US 5658730 A	19-08-1997
		AU 4473596 A	19-07-1996
WO 9743446 A	20-11-1997	AU 2904197 A	05-12-1997